Improving Patient Record Search: A Meta-data based Approach

Iman Amini[1] David Martinez[2] Xiaodong Li[1] Mark Sanderson[1]
[1]RMIT Dept of Computer Science and NICTA, Australia
[2]MedWhat.com and the University of Melbourne, Australia
iman.amini,mark.sanderson,xiaodong.li@rmit.edu.au
david.martinez@nicta.com.au

Abstract

The International Classification of Diseases (ICD) is a type of meta-data found in many Electronic Patient Records. Research to explore the utility of these codes in medical Information Retrieval (IR) applications is new, and many areas of investigation remain, including the question of how reliable the assignment of the codes has been. This paper proposes two uses of the ICD codes in two different contexts of search: Pseudo Relevance Judgments (PRJ) and Pseudo Relevance Feedback (PRF). We find that our approach to evaluate the TREC challenge runs using simulated relevance judgments has a positive correlation with the TREC official results, and our proposed technique for performing PRF based on the ICD codes significantly outperforms a traditional PRF approach. The results are found to be consistent over the two years of queries from the TREC Medical test collection.

Keywords: Information Storage and Retrieval, Information Search and Retrieval,ICD classification, Pseudo Relevance Feedback

1. Introduction

Electronic Patient Records (EPR) also referred to as Electronic Health Records (EHR) are the digitally stored medical notes written by doctors and practitioners during patients' visits. The volume of EPR is ever increasing. Demand for their use beyond simply recording the health of a person is changing. One potential new application is to use a collection of such records as a source for finding patients for a medical trial. For this task it is necessary

to search over large numbers of EPRs to find patients matching certain criteria, such as suffering a given disease, or belonging to a demographic group. However, because of the structure and vocabulary of the records, search over such content presents new research challenges. In order to start exploring this problem, TREC[1] (Text REtrieval Conference) organized medical IR tracks in 2011 and 2012, where the goal was to identify patient records that fulfill the characteristics of given queries (e.g. "Patients with hearing loss"). The queries were built by targeting a list of research areas that the U.S. Institute of medicine has considered priorities for comparative effectiveness research. Participation in these tracks was strong with 54 research groups submitting runs over the two years it lasted.

One of the differentiating characteristics of search over EPRs is that the target documents contain associated meta-data, such as codes belonging to the International Classification of Diseases (ICD), which are manually assigned to each report by health administration workers.

The ICD is a system for classification of health care, providing a system of diagnostic codes with a large diversity of symptoms, signs and medical findings. The codes are used to help with health informatics processes such as billing for health insurance reimbursement [1].

Other usages of ICD codes are to help with statistics related to the general health of a country, monitor the prevalence of diseases and to be used for the compilation of the national mortality and morbidity statistics. ICD codes have been shown to have problems of completeness and bias [2], and this could harm IR effectiveness. The codes are also challenging to work with, as they have a hierarchical structure with different levels of specificity. For instance *hearing loss* can be linked to many ICD codes, including but not limited to 389.03 (middle ear), 389.0 (conductive hearing loss), and 380.01 (external hearing loss).

Exploiting the presence of the ICD codes in records has not been extensively explored for IR, and our focus here is to enhance ranking methods for medical IR by relying on those codes. The main reason for examining their use, is that queries in the domain tend to refer to diseases, and the ICD codes carry information summarizing the patient's diseases and health conditions.

Since the workers who assign the codes to EPRs are required to follow strict guidelines, the use of ICD codes could help to alleviate some of the

[1] http://trec.nist.gov

imprecision present in a bag-of-words representation of such records. This is particularly important in patient records, as often the free text part of the record will contain speculation, negations (e.g. "the patient does not have X"), references to past conditions, family history of the patient, etc. ICD codes, on the other hand, refer to the current conditions of the patient.

Apart from being used to enhance retrieval effectiveness, they have also been studied as a source of evidence for building test collections for medical IR [3]. Here researchers have speculated that ICD codes can accurately summarize the content of queries and documents, and can be used as a proxy for relevance judgments (*qrels*) in IR test collections. However, a limitation of this previous work is that assessments have never been compared to real clinical queries.

In this paper, we exploit ICD codes for medical IR in two ways. We perform the first systematic analysis of the use of ICD codes for pseudo-relevance judgment (PRJ), by comparing the ranking of runs submitted to TREC based on qrels and the ranking based on qrels built from ICD codes. We use runs submitted to the first Medical TREC track [4] in 2011 (*TREC-M1*), and the second track [5] in 2012 (*TREC-M2*). This analysis intends to address the following research question: *How reliable are the ICD codes for automatic judgment of medical records?*

For our second contribution, we explore whether the direct approach of simply mapping ICDs into their textual representation is the most effective way of using this resource. To answer this question we introduce a novel IR method that relies on ICD codes for a form of pseudo-relevance feedback (PRF), and we compare search based on this system with a common approach for exploiting ICD codes.

The rest of the paper is organized as follows. We present the prior work related to this research with a short survey on techniques used in medical IR, and also previous approaches for PRJ. Next, we describe the test collection and methodology to build our PRJ framework, followed by our proposed approach to model the ICD based PRF. The paper concludes with the analysis of the results, discussion of the limitations, and the future work.

2. Background

The ability to conduct research on the retrieval of clinical records has been limited in previous years, due to the lack of a publicly available dataset of appropriate size. The bulk of the work in this area has been focused on

Natural Language Processing challenges, such as extracting specific information from a small number of clinical records [6], while the IR research on biomedical text has focused on searching the literature. However in 2011 the TREC medical retrieval track was introduced, and this generated much interest in the IR challenges of search over EPRs [4, 5]. We describe here the main medical IR approaches that are related to our work, as well as previous work on PRJ, which is the other research focus of this paper.

2.1. Information Retrieval for Patient Records

The TREC Medical Challenges of 2011 and 2012 give the best picture of the state of the art at this point, since several research groups participated on a shared task over the same patient repository. The best runs in 2011 [7] and 2012 [8] focused on different aspects of the search. King et al. in 2011 relied heavily on text processing and information extraction. They tuned their system using their own manually created relevance judgment of approximately 190 reports per query. In 2012 Zhu and Carterette relied on evidence aggregation, external query expansion and Markov Random Fields. They employed 3 levels for merging the results of IR systems by evaluating visits, based on the best evidence from the reports, aggregation of reports to a visit, and finally the combination of both approaches. The 2012 winning system benefited from the availability of training data from 2011, and they performed optimization of parameters over the early query set. Both groups gained improvement by using external knowledge sources for query expansion, however, many other configurations contributed to the performance of their final systems.

Our group participated in both editions of the challenge. In TREC-M1 we mainly focused on external knowledge sources, such as the UMLS, and DBpedia for query expansion [9]. In 2012 we took a different approach by locally expanding the queries with the collection (using pseudo-relevance feedback based on ICD codes), and by detecting and modifying the negated text in the reports [10]. Our 2012 submission is the basis of this article, and here we extend the system by exploring the use of ICD as pseudo relevance judgments, present diverse ways of mapping queries into ICD codes, and evaluate its performance systematically over the 2011 and 2012 medical TREC collections.

The retrieval tasks in both the 2011 and 2012 Medical TREC tracks highlighted that vocabulary mismatch is one of the key problems for the domain. A common way to alleviate the problem is to use some external

resource, such as a biomedical knowledge base or a catalog of terminologies (e.g. ICD codes). The utility of ICD codes is illustrated through the use of this source of information by most participants in the TREC tracks. The most common approach to their exploitation was to expand the text in the medical records, in an attempt to increase the word overlap with the queries (which have no assigned codes). Each ICD code has a short text description and a simple approach used by a number of the TREC participants was to replace each code with its written description. This method was used in the best automatic runs for TREC-M1 [7] and TREC-M2 [8].

A different approach was implemented by Limsopatham et al. [11] who expanded the text in medical records with ICD descriptions and words taken from Wikipedia pages related to the ICD codes. Their system performed well (in the TREC-M1 run of the track) in terms of Bpref, where a marginal improvement was gained over their baseline. However for other two measures (R-prec and P@10), it was outperformed by the baseline. In Bedrick et al.'s [12] work, ICD codes were assigned to queries using an automated method based on a parser, and no clear improvement was reported in this case.

In order to exploit ICD codes for IR, an important step is the automatic mapping of queries into ICD codes. The previous work on assigning these codes to text fragments has focused on document-level evaluation (e.g. patient records); there is no evaluation at the query level that we are aware of. For patient records, in 2007 Pestian et al. [13] curated a shared task with the goal of fostering research on automatic assignment of ICD codes at document level. They provided training data of approximately 1,000 records with 45 ICD codes which made 94 distinct combinations. Various approaches including negation, machine learning and symbolic processing were used by the top participants. Later Aronson et al. [14] found that combining the evidence from multiple classifiers and a pattern matching algorithm in a stacking setting, is indeed more efficient than any individual classifier. They also observed a consistent improvement by using negation to discard negated text from the records.

Apart from the TREC challenge, more recently CLEF organized a shared task, with the purpose of fostering ways to access health data by lay people, in order to understand their health problem [15]. In 2013 shared task, discharge summaries were used by the organizers to generate queries and the collection was built by a crawl of certified health web pages. 9 groups participated in this task, and overall 48 runs were submitted. Although no run

significantly improved over a PRF baseline created by the organizers, there were 4 runs (from the same team) which outperformed the PRF baseline in terms of their precision at 10. The following year, however, CLEF 2014 retrieval task [16] proved more successful, showing both improved baselines and results, providing a cleaner data collection than 2013. The top three systems took advantage of a language modelling approach and query expansion. The best performance was obtained using the UMLS Metathesarus for concept-based retrieval, and mutual information to identify related terms for query expansion.

2.2. Exploiting Meta-data for Patient IR

Apart from ICD codes, there are other resources that have been used to avoid the vocabulary mismatch between queries and documents in the medical domain. The most widely used public repository of medical terms is known as the UMLS (Unified Medical Language System) Metathesaurus [17], and is maintained by the National Library of Medicine (NLM) for its use in biomedical research. This knowledge base integrates different controlled vocabularies, such as ICD codes, Medical Subject Headings (MeSH), and Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT);[2] which we describe below. An example of the use of UMLS was presented by Jain et al. [18], who propose a framework for symptom-based retrieval of medical records. They use several sources of knowledge, including semantic relationships from the UMLS, and terms suggested by medical experts. Five hypothetical patients were randomly generated, each with 200 nursing assessment notes to form the documents, and 10 queries related to various symptoms were selected by a medical expert. The results showed that the terms suggested by medical experts were the most effective source for query expansion, followed by clinically associated terms from the UMLS. However, the combination of all expanded terms from different sources yielded the highest score.

A resource widely used on its own is SNOMED-CT, which is a subset of UMLS, comprising a collection of medical terms covering diseases, findings, procedures, micro-organisms, substances, etc. Koopman et al.[19] converted all the terms in queries and documents of the medical TREC to SNOMED-CT concepts automatically; such that a single SNOMED-CT concept could

[2]http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html

capture all the terms that were associated to it. This was an attempt to eliminate the need to introduce new terms or to perform any kind of relevance feedback to match more semantically related terms. Their results showed significant improvements using their concept-based method, in comparison to a keyword baseline. Martinez et el. [20] also used UMLS as a graph to find concepts for expansion, with promising results over the same TREC collections.

MeSH is a controlled vocabulary thesaurus of the National Library of Medicine[3]. The utility of applying MeSH to medical IR has also been extensively studied, although not specifically for patient records. Lu et al. [21] automatically mapped query words to MeSH terms to extend the original queries. This work was conducted on collections from two TREC genomics tracks (2006,2007) [22, 23]. Lu et al found that their replication of the PubMed's [4] Automatic Term Mapping (ATM) to MeSH terms was effective in finding more relevant documents, while, it did not improve the precision in the top retrieved documents. Jalali and Borujerdi [24] used the synonyms of the identified MeSH terms or direct descendants of them in the queries as expansion terms, and reported improvement over the various retrieval systems such as expansion with general purpose ontologies.

Mapping medical text (found in queries or in documents) to the concepts and codes described here, is commonly done through the tool *MetaMap* [25]. MetaMap is extensively used in the biomedical text processing and information retrieval communities to map medical terms in text into the concepts held in the (UMLS) Metathesaurus. For example, for the following text,

Cancer patients with liver metastasis treated in the hospital who underwent a procedure

MetaMap identifies these phrases: "liver metastasis", "treated hospital", "cancer patient", and "procedure".

The Metathesaurus concepts identified in queries have been used as candidate keywords for query expansion, or applied to stop non-informative words from queries [10]. Each concept in the Metathesaurus is assigned one or many Semantic Types (STs) which relate to one another through the semantic networks in the UMLS ontology [17, Chapter 5]. To avoid the risk of

[3]http://www.nlm.nih.gov/mesh

[4]http://www.ncbi.nlm.nih.gov/pubmed

expanding non-informative terms, the STs can be used to limit the selection of candidate terms only to those associated to a subset of STs. Karimi et al. [26] manually identified two STs to safely select the terms for expansion ("Pharmacologic Substance" and "Therapeutic or Preventive Procedure"), with which they achieved competitive results in Medical TREC 2011.

2.3. IR Evaluation with Pseudo Relevance Judgments

Generating relevance judgments is always time consuming and due to limitations on human resources, the TREC community conventionally assesses only a portion of the collection; usually by incorporating the top k documents from all the runs submitted by the participants. This is known as *depth-k pooling*. Shallow pooling techniques are those where the average number of judged documents per query is low. In the TREC 2011 medical track, shallow pooling was employed, and the cost of obtaining manual judgments was highlighted by the track organizers. Since shallow pooling makes evaluation less reliable, means of alleviating the problem of obtaining judgments are always being sought.

Soboroff et al. [27] produced a widely cited paper, which described an attempt to automatically build relevance judgments by using patterns of occurrence of documents retrieved by multiple IR systems. The results appeared promising although there were limits to the quality of the judgments, particularly for measuring highly effective retrieval systems. A conclusion that can be drawn from Soboroff et al.'s work is that some level of manual intervention is required when forming relevance judgments.

More recent work on pseudo-relevance judgment [28] has shown that by relying on documents retrieved frequently by a diverse set of systems, it is possible to build relevance judgments automatically, and achieve high correlation with manually judged data. They concluded that the simple method which was based on the ordering of the documents in the pool by the number of runs that returned each document at or above rank 30, performed as well as any other existing system. Their method yielded higher correlation on the NTCIR collection than TREC. Nonetheless, all the past work relied on the pool of runs gathered from the participants in the shared tasks, which is not an option at the time of building a test collection.

Building on ideas tried before, where category structure was used as a substitute for relevance judgments (e.g. [29]) Koopman et al. [3] treated the written descriptions of ICD codes as queries and the documents containing those codes as the relevant set for that query. However, the queries were

artificially created, and it is not clear whether they would be representative of real world questions formulated by medical professionals. In addition, the quality of these wholly simulated judgments could not be compared to the manual judgments produced for TREC, as they used different queries.

While other attempts revolved around simulating the entire qrel, Mollá et al. [30], assuming relevant documents bear some degree of resemblance, automatically completed partial and limited qrels, by gathering unjudged documents that are similar to the judged relevant documents. They reported positive correlation when the number of available qrels is very limited.

3. Material and Methods

This section details the test collection and methods of evaluation used in this paper, followed by the two aspects of medical search investigated in this paper, PRJ to alleviate the cost of IR evaluation and PRF to improve search effectiveness.

3.1. Test Collections

To perform our experiments, we used the data set from the last two Medical tracks of TREC for all the experiments. The two collections shared the same set of documents, which are medical records collected in 2007 during the course of one month, from multiple hospitals in the U.S.

The records consist of 93,552 clinical *reports* of patients visiting departments within hospitals. A patient could *visit* multiple departments during his/her time at the hospital. The TREC organizers provided a one-to-many mapping table from multiple reports to single visits. There are 17,265 visits in the collection. Nearly one fifth of the visits consist of a single report; the rest have multiple reports ranging from two to one hundred.

Each report contains four informative XML tags: Two tags are reserved for the assignment of the ICD codes, namely, *Admit Diagnosis* and *Discharge Diagnosis*; a third tag is a short text (truncated to 40 characters) naming the chief complaint; and the main body of the text is given in a separate tag. Although the main text of the reports is not systematically structured; it has headings, which represent the start of a new section. We refer to these sections as fields, and the following are some instances of them: *Family History*, *Present Illness*, *Allergies*.

In the first year of the medical track there were 34 queries, and 47 in the second year. The queries were built by targeting a list of research areas

that the U.S. Institute of medicine has considered priorities for comparative effectiveness research[5], and the relevance judgment was done by groups of clinicians after pooling documents for each query. The queries included different pathologies and treatments, as well as demographic constraints. Four queries were removed due to the insufficient number of relevant visits (i.e. documents) in the collection.

3.2. Evaluation Paradigm

The evaluation in this paper is two-fold: correlation analysis followed by measurement of search effectiveness.

Evaluating the performance of the PRJ involves testing how similar the pseudo rankings are to the ranking produced by the official relevance judgment. Kendall's tau has been widely used in IR for such purposes. It measures the number of pairwise swaps between the rankings until the two are the same, and is normalized in a way that it produces 1.0 if the ranking are the same and -1 where the rankings correlate reversely. Equation 1 is the formula for Kendall's τ that we used in this paper, which is the proportion of concordance pairs C versus the proportion of discordance pairs D, where n is the total number of pairs.

$$\tau = \frac{C + D}{\frac{1}{2}n(n-1)} \tag{1}$$

On the other hand comparison between IR systems has been historically based on two different aspects of search results, precision and recall. Most of the popular evaluation metrics such Mean Average Precision (MAP) are a combination of the two. However, one limitation of MAP is that it assumes that the relevance judgment is complete, which is not true for most cases. Therefore in 2004 an alternative metric known as Bpref was proposed, whereby the differences between systems are measured based on the number of judged not relevant documents prior to the relevant documents. Bpref was chosen to be the main evaluation metric for TREC-M1 and we also report all the system effectiveness scores in terms of Bpref. The formula for Bpref is given in Equation 2, where R is the number of relevant document for a

[5] `http://www.iom.edu/Reports/2009/ComparativeEffectivenessResearchPriorities.aspx`

query, r is a relevant document and n is a member of the first R judged non-relevant documents retrieved by the system.

$$Bpref = \frac{1}{R} \sum_{r} \frac{|n \, ranked \, higher \, than \, r|}{R} \tag{2}$$

3.3. Pseudo Relevance Judgment derived from ICD Codes

For our first experiment, we tested whether relevance judgments for a medical test collection can be derived from ICD codes. First we assigned the codes to each of the queries in TREC-M1 and M2, by relying on both manual and automatic processes. Documents containing at least one of these codes were assumed to be relevant for the given query (pseudo-relevant from now on). We tested this approach, by comparing the way that the two forms of qrels (the real ones and pseudo-qrels) ranked retrieval systems against each other; a correlation-based method that is widely applied in the IR literature (e.g. Büttcher et al. [31]).

Throughout this section we refer to two versions of relevance judgements, namely pseudo and real qrels. Real qrels are the relevance judgements provided by the TREC challenge organisers, and they are built by manually assessing documents. Pseudo qrels are built automatically for each query, by associating ICD codes to the query, and then considering all the documents in the collection which have at least one of those codes as relevant for the query.

3.3.1. Assigning ICD Codes to Queries

The problem of automatically assigning ICD codes to medical records has been tackled as a classification problem [13], where a number of training instances were used in a shared task to develop machine learning classifiers to predict the ICD codes in the test data. However we do not have training data for classifying medical queries into ICD codes, and we treat this as an unsupervised problem for which we develop three automatic approaches, that we call ICD coders.

First we manually map TREC queries into sets of ICD codes. Every code has a description which we extracted from the ICD9Data web site[6]; see Table 2 for examples of the codes. Note that, the examples are only of

[6]http://www.icd9data.com/

type "disease", as the presence of other kinds of ICD codes such as "procedure" was indeed limited in the TREC collections. The manual look up for ICD codes involved two of the authors querying all the disease names found in the TREC queries, in order to locate the best ICD matches. Each author performed this task separately, and the gold standard was built after discussion of each disagreement. Some of the queries contained different boolean operators linking the diseases (e.g. "patients with AIDS who develop pancytopenia"), and the way to represent these was the main source of disagreement. We decided to represent the queries with ICD-codes linked via boolean operators.

In order to measure the level of agreement we calculate the overall percentage of the overlap between the ICD assignments per query. The percentages of agreement for TREC-M1 and TREC-M2 were %55 and %44 respectively. An instance of a query with disagreement is the following. For the query "Patients who underwent minimally invasive abdominal surgery", one author did not assign any ICD code and the other author assigned, 789.0,550,553, assuming that if the patient had surgery, she must have had abdominal pain. It was decided not to assign any codes for cases where the matching is an approximation, therefore the above query did not receive any ICD code in the gold standard. After the first annotation phase, the authors discussed all the disagreements, and reached joint decisions for each case. We used this final set as the gold standard.

In order to measure the level of agreement we calculate the overall percentage of the overlap between the ICD assignments per query. The agreement considers whether the annotators assign exactly the same codes, and we report the percentage of overlap between the ICD assignments of the two annotators.

The percentages of agreement for TREC-M1 and TREC-M2 were %55 and %44 respectively. In order to be unanimous in our ICD assignments, we decided not to assign any codes for cases where the matching is an approximation, in other words, we only assigned ICD codes for diseases that were explicitly mentioned and not implied by the intervention.

After the first annotation phase, the authors discussed all the disagreements, and reached joint decisions for each case. We used this final set as the gold standard.

Assigning ICD codes to the short TREC queries using the online resources was rather a straightforward task. Albeit to ensure the quality of our ICD assignment we gathered all the ICD codes from the relevant documents per

Query-Number	Query
113	Adult patients who received colonoscopies during admission which revealed adenocarcinoma
122	Patients who received total parenteral nutrition while in the hospital
137	Patients with inflammatory disorders receiving TNF-inhibitor treatments
146	Patients treated for post-partum problems including depression, hypercoagulability or cardiomyopathy
174	Elderly patients with ventilator-associated pneumonia
183	Patients presenting to the emergency room with acute vision loss

Table 1: Six queries that do not have a matching ICD codes in the collection

query basis and compared them against our manually assigned ICD codes. The intersection of the ICD codes in the relevant documents turned out to be null for most of the queries. Hence the gold standard for each query is the union of the ICD codes found in the documents that are judged relevant for a particular query.

With the exception of six queries given in Table 1 all other queries aligned with our manually assigned ICD codes against the relevant documents. That is, the entirety or a subset of the manually assigned ICD codes were found in the gold standard.

On aggregate across the whole set of queries there has been 75.46% alignment between the ICD codes from the gold standards, and the set of ICD codes assigned by the authors. This breaks down to 3 set of queries, some with no alignment (6 queries), and the other two sets with either 100% (50 queries) or less than 100% alignment (25 queries).

We verified the accuracy of our ICD code assignment, however, we learned that the TREC collection lacks a number of procedural and disease related ICD codes, which is the case for the aforementioned six queries. Furthermore, in alignment with our finding, Bedrick et al. [12] mention that, as an artifact of the TREC data export process, the number of ICD codes per visits may have been truncated to a certain number, which indicates the possible loss of important ICD codes in some records.

We then explored automatic mapping of queries into ICD codes. We developed three ICD coders for the task, each using different resources and means of matching query and the text in target codes. The first coder relied on word overlap between the ICD description text and query terms. An IR system was used to find the ICD description that best matched the query. The system was configured to use a PL2 [32] weighting model. The terms in the queries and ICD descriptions were stemmed using Porter stemmer [33], and a stop word list from Goodwin et al. [34] was used. This method assigns

V58.66	Long-term (current) use of aspirin
596.7	Hemorrhage into bladder wall
585.6	End stage renal disease
786.8	Hiccough
941.13	Erythema due to burn (first degree) of lip(s)
783.40	Unspecified lack of normal physiological development
V44.4	Status of other artificial opening of gastrointestinal tract
952.08	C5-c7 level with central cord syndrome

Table 2: A sample of ICD codes with descriptions

a single ICD code to each query.

Our second approach used the information boxes in Wikipedia to obtain the ICD codes of the concepts in the query. The process follows the following steps: (i) apply MetaMap to identify the set of medical concepts in the queries, (ii) automatically retrieve the Wikipedia page of each concept, and (iii) extract the ICD codes found in the information box for each of the retrieved pages.

The editors of Wikipedia often include redirects to a medical term from synonym terms. This means that by searching for any of the variant forms of a given disease, Wikipedia will return the main page describing the concept. For example, Wikipedia does not have a page match for the term *hearing loss*, however an attempt to look up such a page automatically redirects to the page about *deafness*[7], which provides appropriate ICD codes. In this case each of the phrases identified in the query can provide ICD codes, and we assign all of them to the query. The maximum number of ICD codes assigned to a query by this coder in TREC-M1 and TREC-M2 were 3 and 6, and the averages were 1.45 and 1.95, respectively.

While the second approach uses the MetaMap indirectly to look up for ICD codes, in the third approach MetaMap is used directly to query UMLS concepts from a UMLS table called *MRCONSO*. The queries are first mapped into UMLS concepts, and the concepts are then used for querying ICD codes.

We illustrate the process with the example query "patients with AIDS who develop pancytopenia". In this case, MetaMap recognizes the terms *Patient*, *Developing*, *AIDS* and *pancytopenia*, that is, four UMLS mappings

[7]http://en.wikipedia.org/wiki/Hearing_Loss http://en.wikipedia.org/wiki/Deafness

Method	Average	Min	Max
Manual	1.44	0	6
MetaMap	0.77	0	4
Wikipedia	1.08	0	6
Ranked	1.00	1	1

Table 3: Average, minimum and maximum number of ICD assigned to each query using different methods, for TREC M1 and M2 combined

are assigned to the query. The Wikipedia coder uses each of these four terms separately and extracts the ICD codes if they are present in the wiki page. In case of the MetaMap coder, four SQL queries are submitted to the *MRCONSO* table to find the relevant ICD codes. The mapping for the Ranked coder is straightforward, and the entire query is used to search over the dataset containing all the ICD descriptions.

In this case we assigned 042,284.1 manually in the gold standard and the Wikipedia coder assigned 284.1,042,044, Ranked method, 248.1, and the direct approach using MetaMap did not find a match.

Table 3 provides further statistics on the number of ICD codes assigned to each queries. Note that the Ranked-based coder always assigns 1 ICD code per query as it is developed. The mapping of queries will be made available from the authors on request.

We compared the performance of our three ICD coders based on precision and recall using the manually assigned ICD codes as a gold standard. Since the coders do not parse the query for boolean operators, and simply return one or more codes, we evaluated them as a multiclass classification problem. However, if the ICD code for a term is within a range, such as the code *140-239* for *Cancer*, we count them as one code, for the sake of evaluation.

Table 4 shows the performance for each of the techniques. The scores are low, but they suggest that the Wikipedia coder is more reliable for our next experiment. We will explore in Section 3.4 whether the low query-mapping performance can still lead to improved IR results.

Next section investigates a different aspect of the ICD codes. We test the usability of the ICD codes to enhance the effectiveness of IR systems.

3.4. ICD based Query Expansion

We describe here the baselines systems followed by the PRF, our ICD based query expansion method.

	Precision	Recall	F1-measure
Ranked	0.39	0.32	0.35
Wikipedia	0.66	0.50	0.57
MetaMap	0.44	0.33	0.37

Table 4: Evaluation scores for the three automatic ICD coders on the combined set of TREC M1 and M2

3.4.1. Baseline Systems

As our main baseline system, we apply the Inverse Expected Document Frequency model with Bernoulli after-effect and the normalization two from the DFR framework which is available in the Terrier [35] open source search engine. This model was chosen as it was found to be the best performing ranking model out of all the available ranking models in the Terrier package, during our first TREC participation [9]. This was important as we wanted to see whether our technique can improve an already robust baseline. The DFR models are instantiated by three components of the framework: selecting a basic randomness model i.e., Inverse Expected Document Frequency in this case, applying the first normalization, and normalizing the term frequencies, for which we use the second normalization form given in the Equation 3.

$$tfn = tf \cdot log(1 + c \cdot \frac{sl}{dl}) \tag{3}$$

Where tfn is the normalized term frequency, tf is the term frequency of t in the document, sl is the standard length and dl is the document length and c is the hyper parameter whose value was fixed to 1, which is the default setting.

Our baseline system incorporated negation detection using the NegEx algorithm with the rule-based Neg-Aggressive setting [10] whereby a concept and its further occurrences are prefixed with the string "no", if the concept is found to be negated at least once within the same report. Altering the word prevents it from being matched when retrieving a positive query term, but it allows to find it in cases where the query is negated (e.g. "Patients taking atypical antipsychotics without a diagnosis of schizophrenia").

Documents were stemmed using the Porter stemmer, and both queries and documents were filtered using the stop-word list recommended by King et al. [7]. All the ICD codes were expanded in the documents replacing the code by the textual descriptions. This was the most common approach to

System	TREC-M1	TREC-M2
TREC-Mean	0.404	0.305
TREC-Median	0.427	0.328
TREC-Best	0.552†	0.451†
ICD-naïve	0.509†	0.330
Traditional-PRF	0.530†	0.343

Table 5: Evaluation of baseline and TREC Best Automatic run and TREC Mean: † indicates significance $p < 0.01$ compared to the TREC-Mean

exploit the ICD codes over the last two medical TREC competitions.

We refer to our main baseline as ICD-naïve, because the only usage of the ICD codes is via mapping of the numeric representation into the text it refers to. In addition to ICD-naïve, we implement the pseudo relevance feedback system from the Terrier package, to see how our modified PRF performs in comparison. We refer to this system as Traditional-PRF. This system has the same setting as the ICD based PRF, described in section 3.4.2. The key difference, however, is that the Traditional-PRF uses the text representation of the queries to select the top documents for expansion, rather than the ICD codes.

For better comparison, we measured the effectiveness of the systems against the mean Bpref scores of all submitted runs to TREC. Table 5 shows that our baselines are already above the TREC mean for both years of the TREC track. The difference is statistically significant for TREC-M1, but not TREC-M2, suggesting that more sophisticated systems competed in the second edition. For TREC-M1 our baseline is close to the winning participant. Note that we employ a 2 tailed student t-test for all the statistical differences reported in this paper.

3.4.2. Pseudo Relevance Feedback Using Evidence from the ICD Codes

PRF is a type of query expansion that identifies salient terms from the local collection and adds them to the initial query based on the the assumption that top n retrieved documents are indeed relevant. PRF is therefore two fold, running the initial query and then using the retrieval results to expand and re-run the query. However, we build a variation of the traditional PRF based on the mapping of the query terms into ICD codes.

We illustrate our query expansion IR approach in Figure 1. First, the Wikipedia-based ICD coder is used to map a query into one or more ICD

code(s). These codes are passed to a 'Document Selector' that gathers relevant reports containing at least one of the codes assigned to the query. We then pick the best report per visit by ranking each report separately against the queries. Preliminary experiments with our training data showed that adding terms from all the reports was rather detrimental. The explanation for which is, adding terms from all the reports can potentially introduce noise as the reports belonging to a visit have the same ICD codes, but do not necessarily have the same context.[8] Therefore one report will be more relevant to a query than others. We found that choosing the most relevant report is a safer approach. Reports were ranked using the Inexp_B2 ranking function in Terrier against the original query, and the scores were used to determine the best report per visit. We assumed that if a report did not appear in the ranking list, it was not relevant and hence removed from the PRF. However, there is no limit to the number of visits used to expand each query. However, only the top ranked n terms will be selected to expand each query, such that we avoid over-expansion problem. The minimum,maximum and mean of visits used for query expansion across two TREC collections are, 0, 91.02 and 611 respectively. Note that, in terms of efficiency when the number of records are much higher we may have to limit our search to the top n documents.

Note that each visit in TREC collections maps into multiple reports and not all the reports within a visit relate to one query necessarily, although they all have the same ICD codes.

All the terms in the reports chosen from PRF were weighted using their normalized term frequency by BoseEinstein-1 [32], the DFR model for expansion. Finally we selected the default top 40 terms for expansion. We refer to this method as *ICD-PRF*. We chose to select the default setting of our search engine to add the top forty terms and did not exhaustively tune all the ranking parameters, since the parameter space could be extremely large, and we did not intend to solve an optimisation problem, but rather trying to see the effect of ICD-codes in certain conditions.

[8]In the case of PRF we want to select the most informative words to expand a query for which we need to find the most relevant report (best report) within a given visit. However, note that the unit of retrieval used for the PRJ experiment is a visit, where a visit may map to multiple reports. Such that, regardless of which report is more relevant to a query for a given visit, we can only use a visit-id in the qrel, and hence the notion of the best report is not relevant for PRJ.

4. Results

In this section we provide details on the results of PRJ followed by the PRF system.

4.1. PRJ Evaluation

For this experiment, we rely on the assignment of ICD codes to queries (manual or automatic) to build our pseudo-relevance judgments. Once we assign the codes to the query, we consider all documents carrying at least one of the codes as relevant (and the rest as non-relevant). After building the pseudo relevance judgments in this manner, we collect all the runs submitted to the TREC-M1 and TREC-M2 evaluations (downloadable from the TREC web site) and use the same set of 34 and 47 queries to rank all the systems based on the Bpref scores obtained by using the official qrels and the qrels from the PRJs. We rely on Bpref because it was the metric of choice for TREC-M1; it was selected because of its robustness for incomplete judgments sets, since it is computed on the basis of judged documents only [36]. Bpref is inversely related to the fraction of judged non-relevant documents that are retrieved before judged relevant documents. The inferred metrics chosen for TREC-M2 had stability problems when applied in TREC-M1.

Our next step is to measure the Kendall τ correlation between the ranking of runs based on the PRJs (from manual and automatic ICD code assignments), with the ranking of runs based on the original TREC manual judgments. The values of τ are shown in the top two rows of Table 6. We find positive correlations in all cases, and the results are similar to previous pseudo-relevance correlation scores [27], where τ correlation numbers were reported to range from 0.369 to 0.571. The correlation was much higher for TREC-M2 than for TREC-M1, and this suggests that the PRJ would be more appropriate for the queries in TREC-M2 (both with manual and automatic assignment). For TREC-M1 the correlation scores are low, and surprisingly the use of manual ICD codes performs slightly worse than the fully automatic system.

We tracked one possible reason to the differences between collections, by analyzing the depth of the pool (section 2.3) for relevance judgments in TREC-M1 and TREC-M2. The average size of judgment set for TREC-M2 (512) was almost twice the size of TREC-M1 (260.7) and the number of judged runs were much higher in the case of TREC-M2 (88) than TREC-M1

	TREC-M1	TREC-M2
PRJ-Manual	0.35	0.59
PRJ-Automatic	0.37	0.50
Max(Official Query Split)	0.41	0.53

Table 6: τ Correlations between PRJ and official relevance judgement

(47). This indicates that TREC-M2 should offer a more robust evaluation framework.

In order to put the correlation scores in perspective, we then calculated the correlation of the system rankings, when measured by different splits of the queries using the official qrels. This is referred to as *Data-based reliability indicator* and has been used to measure test collection reliability in the past [37]. The intuition behind this experiment is that the rankings resulting from the different subsets of queries should present a reasonably high correlation, given that the queries and relevance judgments originate from the same source, and they are manually assigned. Since we can only use half of the queries in each collection, we report here the highest correlations that we obtain, in order to compensate for having smaller query sets. For each collection (TREC-M1 and TREC-M2), the queries were randomly divided into two equal subsets. The random division of the queries was performed for 1,000 iterations and the average Bpref score for each run in each of the subsets was calculated based on the official qrels. For TREC-M2, to make the number of queries equal in each subset , we eliminated the last query, i.e., query 182. Next, we calculated the Kendall τ correlation for every iteration, and obtained 1,000 correlations for each collection. Finally we selected the highest correlation score for TREC-M1 and TREC-M2. The Kendall τ for this reference approach is given in the bottom row of Table 6, named "Max(Official Query Split)". We can see that the τ scores are comparable to the manual pseudo-relevance for TREC-M1, and for TREC-M2. This indicates that the use of pseudo-relevance judgments performs similarly to the reliance on half of the queries of the collection with real qrels.

Table 7 presents a subset of queries, including their manually-assigned ICD codes and their corresponding τ correlations between the PRJ and the official relevance judgment. This subset represents the four queries with the highest and the lowest correlations, and the descriptions of the codes are given in Table 8. The queries at the top and the bottom do not seem very

Query ID	Query	ICD codes	Tau Correlation
156	Patients with depression on antidepressant medication	(296.2 OR 296.3 OR 311)	0.71
160	Patients with Low Back Pain who had Imaging Studies	724.2	0.685
182	Patients with Ischemic Vascular Disease	410-414	0.681
148	Patients acutely treated for migraine in the emergency department	346	0.669
123	Diabetic patients who received diabetic education in the hospital	250	0.052
133	Patients admitted for care who take herbal products for osteoarthritis	715	0.059
135	Cancer patients with liver metastasis treated in the hospital who underwent a procedure	(155.0 OR 155.2)	0.068
110	Patients being discharged from the hospital on hemodialysis	(584 OR 585)	0.076

Table 7: Four highest and lowest τ correlations between the official and the PRJ using the manually assigned ICD codes

different, with most of the queries having close matches to ICD codes, and containing restrictions that cannot be directly captured with ICD codes (e.g. "Imaging studies"). However the differences in performance suggest that some restrictions have a greater effect in the relevance of documents.

Figure 2 and 3 graph the official TREC Bpref scores on a per query basis, which we compare to the Bpref scores obtained using our manually created qrels for TREC-M1 and TREC-M2 respectively. It can be seen that TREC-M2 is more consistent with the official scores and the top and low systems are mostly similar. Figure 4 and 5 show the same correlation using the Wikipedia-based automated ICD coder, explained in section 3.4 where it seems that the signal is weaker in this case.

Despite the positive correlations, we can see that relying solely on ICD codes does not provide a reliable evaluation framework and this can be taken into account by efforts such as Koopman et al. [3].

ICD code	Description
296.2	Major depressive disorder single episode
296.3	Major depressive disorder recurrent episode
311	Depressive disorder, not elsewhere classified
724.2	Lumbago
410	Acute myocardial infarction
411	Other acute and subacute forms of ischemic heart disease
412	Old myocardial infarction
413	Angina pectoris
414	Other forms of chronic ischemic heart disease
346	Migraine
250	Diabetes mellitus
715	Osteoarthrosis and allied disorders
155.0	Malignant neoplasm of liver, primary
155.2	Malignant neoplasm of liver, not specified as primary or secondary
584	Acute kidney failure
585	Chronic kidney disease (ckd)

Table 8: ICD codes, and their corresponding definitions.

	ICD-naïve	Traditional-PRF	ICD-PRF
TREC-M1	0.509	0.530	0.547
TREC-M2	0.330	0.343	0.374††
TREC-combined	0.406	0.422	0.446†

Table 9: The first two columns present the Bpref scores of our baselines, and are followed by the performances of our ICD-PRF method. The scores are followed by up to two signs. † implies significance ($p \leq 0.01$) difference over the ICD-naïve and †† means significance difference ($p \leq 0.01$) over both ICD-naïve and the Traditional PRF.

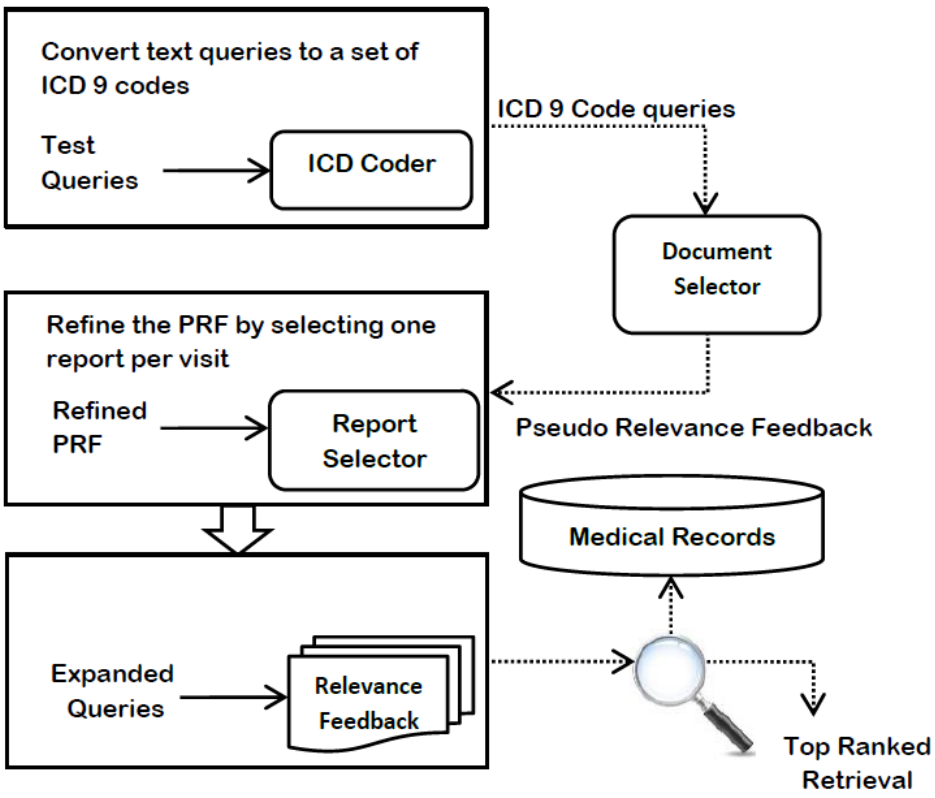

Figure 1: The scheme used for expanding the original queries based on the ICD Pseudo Relevance Feedback

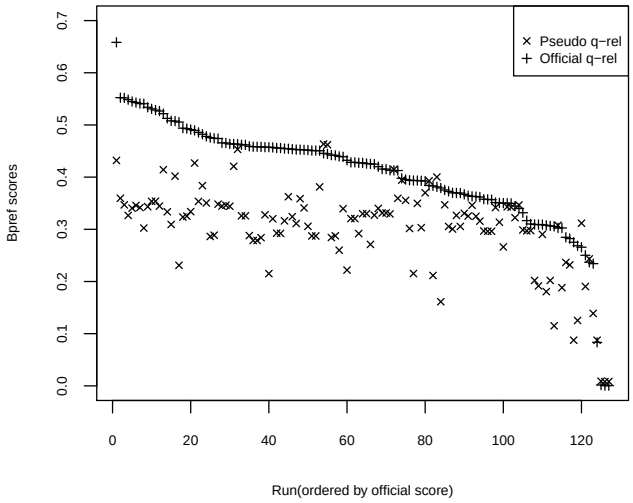

Figure 2: Official Bpref scores vs. ICD based Pseudo Relevance judgment for TREC-M1 with manual ICD assignment

4.2. PRF Evaluation

We tested the systems on three different sets of queries: the 34 queries of TREC-M1, the 41 of TREC-M2 and a combination of both. Table 9 shows the comparison between the two PRF methods and the baseline (ICD-naïve). The improvement of the ICD-PRF method over the ICD-naïve is consistent across the different query sets. However, the level of significance is higher for the larger set of queries when the test queries are combined. The ICD based PRF system yields higher scores than the traditional PRF, and the difference is statistically significant for TREC-M2.

One of the reasons for the good performance of our approach may be that some relevant reports do not have term overlap with the queries. Tinsley et al. [38] found that for one of the TREC queries, only one visit out of all the relevant visits contained a string from the query, while all the relevant visits contained the corresponding ICD codes. We suspect that there are more of such cases across the collections leading to improvement in effectiveness over the Traditional PRF.

We also showed that the significant improvements over the ICD-naïve

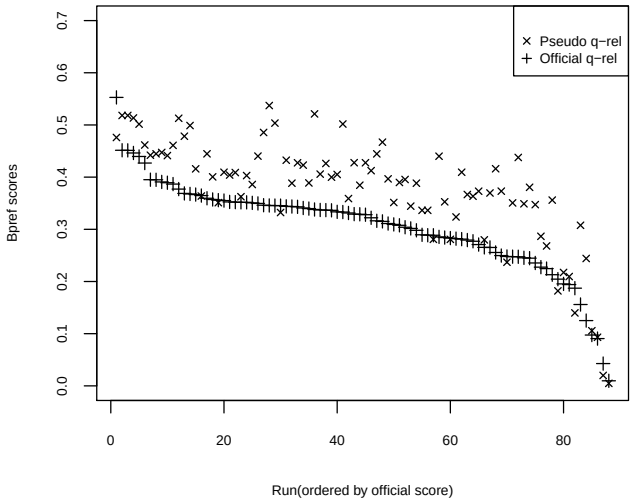

Figure 3: Official Bpref scores vs. ICD based Pseudo Relevance judgment for TREC-M2 with manual ICD assignment

proves that there is something to gain by going beyond the simple mapping of ICD codes into their text representation.

5. Discussion

While our ICD expansion method does not outperform the best run in TREC-M2 and it performs equally to the best run in TREC-M1, we demonstrated that a different way of incorporating the ICD code is indeed superior to the current approach. Due to the complexity and the engineering efforts required to replicate the best runs in TREC and the fact that TREC 2011 best run was tuned with the author's manually created training data, we were unable to reproduce these systems. However, we implemented two strong baselines, both incorporating the ICD implementation of the TREC best runs. We showed that performing PRF based on the ICD codes is more effective than the conventional method (Traditional PRF) that relies on the text representation of the queries.

To discover the strength of ICD expansion, we hypothesized that this method performs favorably for those queries which have a strict alignment

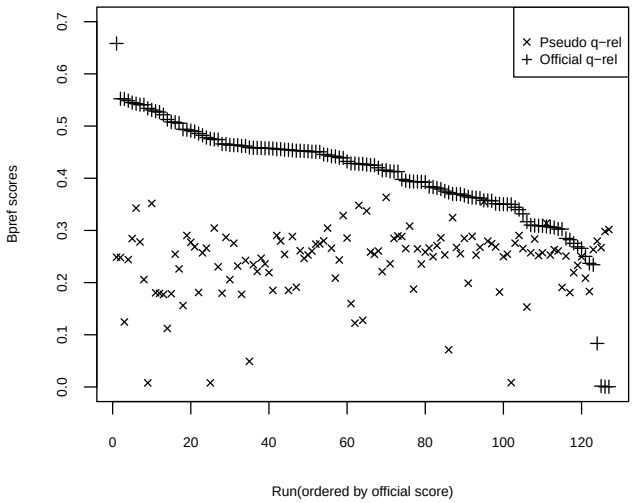

Figure 4: Official Bpref scores vs. ICD based Pseudo Relevance judgment for TREC-M1 with automatic ICD assignment

	ICD-naïve	Traditional-PRF	ICD-PRF
TREC-Combined	0.393	0.386	0.456††

Table 10: Performance of systems for a set of perfectly aligned ICD codes double † implies significance difference ($p \leq 0.01$) over both ICD-naïve and the Traditional PRF.

with the codes where all the medically related terms have at least one ICD code.

For instance query 123 and 133 in Table 8 are instances of relaxed alignment with ICD codes. They contain terms such as *diabetic education and herbal products* which do not have any related ICD codes. On the other hand the query *Patients with hearing loss* is one with strict ICD alignment.

Overall we found 36 queries with strict alignment across two TREC collections. Table 10 shows the result of this experiment. The score shows that the ICD-PRF is significantly superior to other two baselines for these queries.

As queries for finding clinical trials may not always map to a correspond-

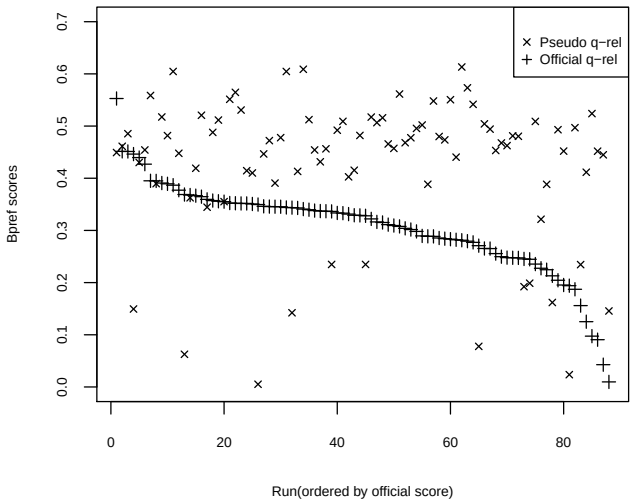

Figure 5: Official Bpref scores vs. ICD based Pseudo Relevance judgment for TREC-M2 with automatic ICD assignment

ing ICD code, an ideal IR system would take advantage of both the traditional and ICD based PRF, where in case of no ICD mapping the former would be utilized.

6. Conclusion

This paper demonstrated the value of a type of meta-data found in most electronic patient records. We believe that this is the first time their use for PRJ over real queries has been conducted. Although we observed positive correlations for TREC runs, we also noticed the difficulty of relying on this sole source of evidence in the scattered plots.

The first part of this paper analyzed the feasibility of a potential alternative to replace the time consuming process of human relevance judgment. ICD codes can designate the primary theme of the medical reports and therefore they were found to be suitable for simulating traditional manual relevance judgments. However, they may not necessarily capture all aspects of the medical queries. In order to incorporate the information about the population or medical devices, if present in the queries, we need to rely on other

sources such as SNOMED-CT concepts. However, no publicly available medical records, to the best of our knowledge, contains such meta-data or the latest version of the ICD codes (ICD 10). We recognize this limitation and await the future shared tasks to provide the latest electronic health records to the community.

Another challenge presented by the queries were *AND* and *OR* conditions found in the formulation of the query. By default our automatic systems used a logical OR operation to gather all the documents with at least one mention of the ICD codes from the queries. We believe that simple rule-based regular expressions to identify and apply such conditions can yield further improvement, although this was beyond the scope of this paper.

Our results in the PRF experiment were more positive, and we measured significant improvements over the traditional approach. We concluded that our approach to using the ICD codes is indeed superior to the simple mapping approach applied by most TREC participants.

The next step towards the refinement of our pseudo judgment and pseudo relevance systems is to automatically designate how relevant a report is against a given query. Human assessors can determine the level of relevance ranging from 0 non-relevant to 2 highly relevant, whereas, our approach indiscriminately assigned 1 to all the relevant visits. In the PRF model this can potentially help in the reduction of the less informative reports. This could also increase the correlation with the official relevance judgment for the PRJ model. While there are ways to compensate for this we left the automatic grading of the relevant documents for future work.

Acknowledgments

NICTA is funded by the Australian government as represented by Department of Broadband, Communication and Digital Economy, and the Australian Research Council through the ICT Centre of Excellence programme.

References

[1] C. Puckett, The Educational Annotation of ICD-9-CM, Channel pub., 2011.

[2] F. S. Roque, P. B. Jensen, H. Schmock, M. Dalgaard, M. Andreatta, T. Hansen, K. Søeby, S. Bredkjær, A. Juul, T. Werge, L. J. Jensen,

S. Brunak, Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts., PLoS computational biology 7 (2011) e1002141.

[3] B. Koopman, P. Bruza, L. Sitbon, M. Lawley, Evaluating Medical Information Retrieval, in: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, ACM, 2011, pp. 1139–1140.

[4] E. Voorhees, R. Tong, Overview of the TREC 2011 Medical Records Track, in: The tenth Text REtrieval Conference, Gaithersburg, MD. National Institute of Standards and Technology, 2011.

[5] E. Voorhees, W. Hersh, Overview of the TREC 2012 Medical Records Track, in: The tenth Text REtrieval Conference, Gaithersburg, MD. National Institute of Standards and Technology, 2012.

[6] E. C. Özlem Uzuner, Imre Solti, Extracting Medication Information from Clinical Text, Journal of the American Medical Informatics Association (2012).

[7] B. King, L. Wang, I. Provalov, J. Zhou, Cengage Learning at TREC 2011 Medical Track, in: Proceedings of TREC, 2011.

[8] D. Zhu, B. Carterette, Exploring Evidence Aggregation Methods and External Expansion Sources for Medical Record Search, in: Proceedings of TREC, 2012.

[9] I. Amini, M. Sanderson, D. Martinez, X. Li, Search for Clinical Records: RMIT at TREC 2011 Medical Track, in: Proceedings of Text Retrieval Conference, 2011.

[10] I. Amini, M. Sanderson, D. Martinez, X. Li, Using Meta-data to search for Clinical Records: RMIT at TREC 2012 Medical Track, in: Proceedings of Text Retrieval Conference, 2012.

[11] N. Limsopatham, C. Macdonald, I. Ounis, G. McDonald, M. Bouamrane, University of Glasgow at Medical Records Track: Experiments with Terrier, in: Proceedings of TREC, 2011.

[12] S. Bedrick, T. Edinger, A. Cohen, W. Hersh, Identifying Patients for Clinical Studies from Electronic Health Records: TREC 2012 Medical Records Track at OHSU, in: Proceedings of Text Retrieval Conference, 2012.

[13] J. Pestian, C. Brew, P. Matykiewicz, D. Hovermale, N. Johnson, K. Cohen, W. Duch, A Shared Task Involving multi-label Classification of Clinical Free Text, in: Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, Association for Computational Linguistics, 2007, pp. 97–104.

[14] A. R. Aronson, O. Bodenreider, D. Demner-fushman, K. W. Fung, V. K. Lee, J. G. Mork, A. Névéol, L. Peters, W. J. Rogers, From Indexing the Biomedical Literature to Coding Clinical Text : Experience with MTI and Machine Learning Approaches, 2007.

[15] L. Goeuriot, G. J. F. Jones, L. Kelly, J. Leveling, A. Hanbury, M. Henning, S. Salanter, G. Zuccon, ShARe / CLEF eHealth Evaluation Lab 2013 , Task 3 : Information Retrieval to Address Patients Questions when Reading Clinical Reports, in: Online Working Notes of CLEF, CLEF (2013), 2013, pp. 1–16.

[16] L. Goeuriot, L. Kelly, W. Li, J. Palotti, P. Pecina, G. Zuccon, A. Hanbury, G. Jones, H. Mueller, ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centred health information retrieval, in: Proceedings of CLEF 2014, 2014, pp. 43–61.

[17] B. (MD), UMLS Reference Manual [Internet], http://www.ncbi.nlm.nih.gov/books/NBK9679/, Sep 2009.

[18] H. Jain, C. Thao, H. Zhao, Enhancing electronic medical record retrieval through semantic query expansion, Information Systems and e-Business Management 10 (2010) 165–181.

[19] B. Koopman, P. Bruza, L. Sitbon, M. Lawley, Towards Semantic Search and Inference in Electronic Medical Records: an Approach Using Concept-based Information Retrieval, in: Proceedings of the First Australian Workshop on Artificial Intelligence in Health 2011, CSIRO Australian e-Health Research Centre, 2011, pp. 1–10.

[20] D. Martinez, A. Otegi, A. Soroa, E. Agirre, Improving search over Electronic Health Records using UMLS-based query expansion through random walks., Journal of biomedical informatics 51 (2014) 100–106.

[21] Z. Lu, W. Kim, W. Wilbur, Evaluation of Query Expansion Using MeSH in PubMed, Information retrieval 12 (2009) 69–80.

[22] W. Hersh, A. Cohen, P. Roberts, H. Rekapalli, TREC 2006 Genomics Track Overview, in: The Fifteenth Text Retrieval Conference, 2006, pp. 52–78.

[23] W. Hersh, A. Cohen, P. Roberts, H. Rekapalli, TREC 2007 Genomics Track Overview, in: The Sixteenth Text Retrieval Conference, 2007.

[24] V. Jalali, M. Borujerdi, The Effect of Using Domain Specific Ontologies in Query Expansion in Medical Field, in: International conference on innovations in information technology (IIT2008), IEEE, 2008, pp. 277–281.

[25] A. R. Aronson, F.-M. Lang, An Overview of MetaMap: Historical Perspective and Recent Advances., Journal of the American Medical Informatics Association : JAMIA 17 (2010) 229–36.

[26] S. Karimi, D. Martinez, S. Ghodke, L. Zhang, H. Suominen, L. Cavedon, Search for Medical Records: NICTA at TREC 2011 Medical Track, in: Proceedings of TREC, 2011.

[27] I. Soboroff, C. Nicholas, P. Cahan, Ranking Retrieval Systems without Relevance Judgments, in: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2001, pp. 66–73.

[28] T. Sakai, C.-y. Lin, Ranking Retrieval Systems without Relevance Assessments - Revisited, in: The Third International Workshop on Evaluating Information Access (EVIA), 2010, pp. 25–33.

[29] V. Harmandas, M. Sanderson, M. Dunlop, Image Retrieval by Hypertext Links, in: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 1997, pp. 296–303.

[30] D. Mollá, I. Amini, D. Martinez, Document Distance for the Automated Expansion of Relevance Judgements for Information Retrieval Evaluation, in: ACM SIGIR Workshop on Gathering Efficient Assessments of Relevance (GEAR), 2014.

[31] S. Büttcher, C. L. A. Clarke, P. C. K. Yeung, I. Soboroff, Reliable Information Retrieval Evaluation with Incomplete and Biased Judgements, in: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, New York, New York, USA, 2007, p. 63. doi:10.1145/1277741.1277755.

[32] G. Amati, Probabilistic Models for Information Retrieval Based on Divergence from Randomness, Ph.D. thesis, University of Glasgow, 2003.

[33] M. F. Porter, An algorithm for suffix stripping, Program: electronic library and information systems 14 (1980) 130–137.

[34] T. Goodwin, B. Rink, K. Roberts, S. Harabagiu, Cohort Shepherd: Discovering Cohort Traits from Hospital Visits, in: Proceedings of TREC, 2011.

[35] C. Macdonald, R. McCreadie, R. Santos, I. Ounis, From Puppy to Maturity: Experiences in Developing Terrier, Open Source Information Retrieval (2012) 60.

[36] C. Buckley, E. Voorhees, Retrieval Evaluation with Incomplete Information, in: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2004, pp. 25–32.

[37] J. Urbano, M. Marrero, D. Martín, On the measurement of Test Collection Reliability, in: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '13, ACM Press, New York, New York, USA, 2013, p. 393.

[38] B. Tinsley, A. Thomas, J. McCarthy, M. Lazarus, Atigeo at TREC 2012 Medical Records Track: ICD-9 Code Description Injection to Enhance Electronic Medical Record Search Accuracy, in: Proceedings of Text Retrieval Conference, 2012.